The vocal effort of dominance in scenario meetings

Marcela Charfuelan, Marc Schröder

DFKI GmbH, Language Technology Lab
Saarbrücken and Berlin, Germany
firstname.lastname@dfki.de

Abstract

In this paper we address two questions about dominance in the AMI-IDIAP scenario meetings: (i) do the annotated most and least dominant utterances correlate with different levels of vocal effort? and if so (ii) how quantitatively discriminative are the vocal effort effects for prosody, voice quality and low level acoustic features? For answering these questions we perform supervised learning with dominance annotations in AMI-IDIAP meetings and vocal effort annotations in controlled data. A linear discriminant analysis (LDA) classifier is used to optimise class separability. We have found that the most and least dominant utterances are acoustically correlated with loud and soft vocal effort. We were able to quantify around 55% discrimination of equal distributions of most dominant, neutral and least dominant utterances using low level acoustic measures.

Index Terms: prosody, voice quality, vocal effort, vocal social signals, acoustic correlates

1. Introduction

In the literature high dominant persons have been characterised by loud, tense voice and low dominance persons by soft, fearful voice [1, 2]. In meetings, dominance has been studied with different purposes and means. For example: in [3] the most dominant person in a group meeting is classified employing speaking length and energy as audio cues, as well as video features; in [2] dominance and role-based status in scenario meetings is predicted using speaking energy and speaking status, along with non-relational and relational cues derived from vocalic and visual cues; in [4] dominance is detected on the basis of easily obtainable features, such as speaking time, number of turns in a meeting, number of words spoken in the whole meeting, etc. In contrast to the previous studies, the present work investigates prosody, voice quality and low level acoustic features of dominance in view of speech synthesis. The long term goal of our investigation is automatic synthesis of expressive speech, so we aim to extract from real data acoustic patterns that can be used to model and re-synthesise different expressions suitable for a range of social signals, like dominance.

In a previous work [5], we have investigated prosody and voice quality of dominance in the AMI-IDIAP scenario meetings and we were able to extract acoustic patterns for two levels of dominance (most and least dominant) and correlate them, to a certain extent, with acoustic patterns of project manager and marketing expert roles. Although at human perception, the project manager role has been found to be the most dominant person and the marketing expert role the least dominant person [2], not always the most and least dominant persons correspond to these roles. Therefore in this work we analyse the voice of the most and least dominant persons irrespective of the role, and extend the study to consider not only prosody and voice quality measures but also low level acoustic features, more suitable to be used in speech synthesis.

Thus in this paper we address two questions about dominance in the AMI-IDIAP scenario meetings: (i) do the annotated most and least dominant utterances, irrespective of the role, correlate with different levels of vocal effort? and if so (ii) how quantitatively discriminative are the vocal effort effects for prosody, voice quality and low level acoustic features? For answering these questions we perform supervised learning with dominance annotations in AMI-IDIAP meetings and vocal effort annotations in controlled data. Sequential floating forward selection (SFFS) and linear discriminant analysis (LDA) are used to optimise class separability and find the best discriminative features [6].

This paper is organised as follows. In Section 2 the databases and methodology is explained. In Section 3 the different acoustic measures are described. In Sections 4 vocal effort is analysed on controlled data to verify the effectiveness of the measures and in Section 5 dominance is analysed using different groups of acoustic measures. In Section 6 conclusions are drawn and future work is envisaged.

2. Data and method

2.1. The NECA vocal effort database

The NECA vocal effort database contains a full German diphone set for each of three levels of vocal effort (“soft”, “modal” and “loud”) and two speakers (one male, one female). Perception as the intended vocal effort was verified using stimuli generated using diphone synthesis voices built from the recordings [7]. Out of the original recordings, 100 short words per speaker and per vocal effort were used in this experiment (600 in total). From this database we have extracted only low level acoustic features (as explained in section 3).

2.2. The AMI meeting corpus

The AMI Meeting Corpus is a multi-modal data set consisting of 100 hours of meeting recordings. Some of these meetings are naturally occurring, and some are elicited, particularly using a scenario in which the participants play different roles. This work focuses on the elicited meetings. In the scenario, four participants play the roles of employees in an electronics company that decides to develop a new type of television remote control. Although the scenario is pre-defined and the roles assigned, the conversations and discussions in the meetings reflect natural interaction [8]. This corpus contains recordings of both video and audio data, orthographic transcriptions and several levels of annotations, for example dominance. The transcriptions include word level segmentation time-aligned to the audio recordings.

The analysis of dominance was performed with 5 AMI meetings held at IDIAP for which dominance annotations are available [3]. 11 sub-meetings (phases), from these five meeting sessions, have been divided into 5 minute segments, and each segment has been annotated by three annotators. The annotators ranked the participation of each person in the segment from highest to lowest, according to their level of perceived dominance. As described in [9], from the annotations, a significant number of the meeting segments, 34, showed full agreement of the most dominant person; full agreement on the least dominant person was found in 31 segments. There were additional segments where 2 out of 3 annotators, the majority, agreed on the most or least dominant person. In this study we only used the full agreement (34+31 segments) annotations.

The distribution of data according to gender, role and dominance of the data used in this study is presented in Table 1. The total number of analysed utterances is 3869. The number of participants is 20, 14 male and 6 female.

Table 1: *AMI-IDIAP meetings: distribution of analysed data (M: male and F: female). The roles of the participants are: Project Manager (PM), Marketing Expert (ME), User Interface designer (UI) and Industrial Designer (ID)*

Role	leastDominant		mostDominant		neutral	
	M	F	M	F	M	F
ID	42	4	0	153	627	102
ME	28	48	34	41	436	101
PM	0	0	516	320	316	106
UI	111	29	0	0	680	175
	181	71	550	514	2059	484

2.3. Method

For each annotated segment of five minutes, we extracted all the participants' utterances and performed acoustic analysis with them. Acoustic measures are extracted for each utterance at frame and utterance level (as explained in section 3). In this work we consider as most dominant utterances all the utterances of the participants rated as most dominant. The same apply for least dominant utterances. The utterances of the participants that were not consider most or least dominant are considered neutral in this work.

Due to the distribution of the data we decided to study the vocal effort effects according to gender. This will not be a problem for our long term objective of synthesis since for synthesis it is always known the gender of the rendered voice. After extracting the acoustic features on the NECA and AMI-IDIAP data, an analysis of variance is performed, to discard features that are not significantly different among the classes. Then the main tendencies on low level acoustic measures on the controlled data are compared with the main tendencies of the same measures extracted from the AMI-IDIAP data. This gave us an idea about the correlation of loud, soft and modal vocal effort with most, least dominant and neutral data. Finally, in order to quantify this correlation, a SFFS-LDA classification is performed on the AMI-IDIAP data. Here we used combinations of prosody, voice quality and low level acoustic features in order to find the set of features that optimise class separability.

3. Acoustic measures

We have selected three sets of acoustic measures: (1) low level acoustic measures are mainly related with measures that can be used directly in speech synthesis. For example, voicing strengths, Fourier magnitudes and Melcepstrum coefficients have been used on implementations of the mixed excitation linear prediction (MELP) vocoder [10]. Other low level acoustic measures give information about the spectrum at sub-band level (spectral entropy) or are articulatory-based. Low level acoustic measures are extracted at frame level, with a frame length of 25 ms. and a frame shift of 5 ms. The frame based measures are averaged per utterance. (2) prosody features are classical features related to pitch, energy, duration, etc. And (3) voice quality measures used mostly in emotion research. Prosody and voice quality measures are extracted at utterance level.

3.1. Low level acoustic measures

- *Voicing strengths:* are estimated with peak normalised cross correlation of the input signal [10]. The correlation coefficient for a delay t is defined by:

$$c_t = \frac{\sum_{n=0}^{N-1} s(n)s(n+t)}{\sqrt{\sum_{n=0}^{N-1} s^2(n) \sum_{n=0}^{N-1} s^2(n+t)}} \tag{1}$$

  In this study one full band (str) and five bandpass voicing strengths are calculated (str1-str5), that is, the input signal is filtered into five frequency bands, with passbands 0-1kHz, 1kHz-2kHz, 2kHz-4kHz, 4kHz-6kHz and 6kHz-8kHz and voicing strengths are calculated for each filtered signal.

- *Pitch harmonics magnitude:* corresponds to the Fourier magnitude of the first ten pitch harmonics of the residual signal obtained by inverse filtering (mag1-mag10) [10]. The Fourier magnitudes capture the shape of the excitation pulse, so it is expected that they capture variations in phonation types.

- *Spectral features*:
  - Melcepstrum coefficients, 25 coefficients (mcep0-mcep24) [11].
  - Spectral entropy, this is a kind of "peakiness" of the spectrum [12]. This value is calculated as follows : the spectrum X is converted into a probability mass function (PMF) normalising it by:

$$x_i = \frac{X_i}{\sum_{i=1}^{N} X_i} \qquad i = 1 \quad to \quad N \tag{2}$$

    where X_i is the energy of the i^{th} frequency component of the spectrum, x is the PMF of the spectrum and N is the number of points in the spectrum. Entropy for each frame is calculated by:

$$H(x) = -\sum_{x \in X} x_i * log_2 x_i \tag{3}$$

    It is expected that spectral entropy is higher for voiced frames than for unvoiced, therefore it has been used in speech endpoint detection and in classification of emotions. In this work spectral entropy is calculated for the full band signal (spec-entropy) and for the five frequency bands filtered signals (spec-entropy1-spec-entropy5).

- *Articulatory-based features*
  - Formants: F_1, F_2, F_3, F_4
  - Formant bandwidths: B_1, B_2, B_3, B_4
  - Formant dispersion:

$$FD = \frac{(F2 - F1) + (F3 - F2) + (F4 - F3)}{3} \tag{4}$$

3.2. Prosody acoustic measures

- Fundamental frequency or pitch ($f0$), extracted with snack [13]
- Pitch entropy (calculated as the spectral entropy)
- maximum, minimum, and range of $f0$
- Duration of the utterance in seconds
- Voicing rate calculated as the number of voiced frames (frames for which $f0 > 0$) per time unit
- Energy, calculated as the short term energy $\sum x^2$

3.3. Voice quality acoustic measures

The following voice quality measures were also used in the study presented in [5]. These measures are based on the calculation of the long term average spectrum (LTAS) in three frequency bands: 0-2kHz, 2-5kHz, 5-8kHz. For each of these bands the maximum level is selected.

- Hamm_effort = LTAS$_{2-5k}$
- Hamm_breathy = (LTAS$_{0-2k}$ - LTAS$_{2-5k}$) - (LTAS$_{2-5k}$ - LTAS$_{5-8k}$)
- Hamm_head = (LTAS$_{0-2k}$ - LTAS$_{5-8k}$)
- Hamm_coarse = (LTAS$_{0-2k}$ - LTAS$_{2-5k}$)
- Hamm_unstable = (LTAS$_{2-5k}$ - LTAS$_{5-8k}$)
- slope_ltas: least squared line fit of LTAS in the log-frequency domain (dB/oct)
- slope_ltas1kz: least squared line fit of LTAS above 1 kHz in the log-frequency domain (dB/oct)
- slope_spectrum1kz: least squared line fit of spectrum above 1 kHz (dB/oct)

4. Testing the measures on controlled data

After extracting low level acoustic features on the NECA database an analysis of variance (one-way ANOVA) was performed on the measures. All the measures were significantly different at 0.1% level (p<0.001), except for Melcepstrum coefficients 3, 9 and 10 for male data and Melcepstrum coefficients 3, 5 and 7 for female data. The LDA 10-fold cross validation classification error for all the data (male and female) was around 3%. This showed that for this "easy" data the low level features capture very well the vocal effort, this is also shown in Figure 1 were the first two components of a principal component analysis (PCA) of low level features is presented. The separation of clusters for female data is more clear than for male data.

In Table 2 the main tendencies of low level acoustic measures for loud and soft vocal effort in the NECA DB are summarised. The arrows in this table show that, for example, the levels of voicing strengths for both male and female data, are increasing on loud voice and decreasing on soft voice. This table shows clear tendencies on both male and female data for Fourier magnitudes, voicing strengths on higher bands and Melcepstrum for higher coefficients.

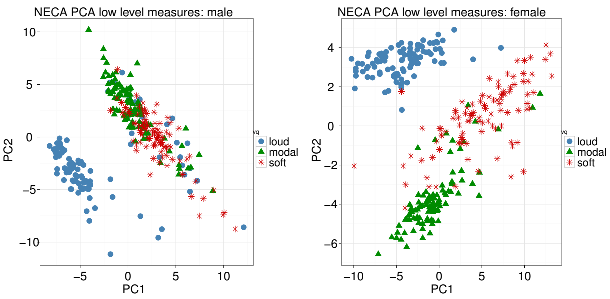

Figure 1: *NECA DB vocal effort discrimination using PCA of low level features. Left: male data (51% of variance explained by the first two PCs). Right: female data (66% of variance explained by the first two PCs).*

Table 2: *NECA DB main tendencies of loud and soft vocal effort according to low level acoustic measures. ⇑ indicates increasing level and ⇓ decreasing level. (The opposite tendency applies for measures not included in this table).*

Measures	Male	Female	loud	soft
Voicing str.	str2-str5	str3-str5	⇑	⇓
Fourier mag.	mag1-mag10	mag1-mag10	⇓	⇑
Formants	F1-F4	F1-F4	⇓	⇑
Formant BW	B1,B4	B1-B4	⇓	⇑
Spec. entropy	entropy 1, 5	entropy 1-5	⇑	⇓
Melcepstrum	mcep 1, 7-24	mcep 3, 5-24	⇑	⇓

5. Acoustic correlates of dominance

After extracting prosody, voice quality and low level acoustic measures from the AMI-IDIAP data an analysis of variance (one-way ANOVA) was also performed on the measures. The following are the measures that were not significantly different (p>0.001) in each set and therefore were not included on the classification experiment:

- Male: F1, F3, B2, B3, spectral_entropy 1, 4, 5, mag 1, 3, 8, mcep 0, 2, 12, 14, 15, 18, 19, 20, 22, 23, 24, max_f0, range_f0, Hamm_breathy, Hamm_coarse.
- Female:str, str1, F1, F2, F3, B1, B2, B3, spectral_entropy 1, 3, 5, mag 2, 4, 6, mcep 0, 3, 4, 8, 9, 11, 14, 19, 23.
- All: F1, F2, spectral_entropy, spectral_entropy 1, mag 2, 3, 5, 6, 8, 9, 10, mcep 2, 5, 8, 9, 10, 18, Hamm_breathy.

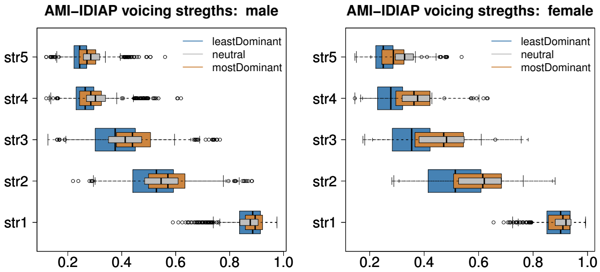

Figure 2: *AMI-IDIAP dominance: boxplot of voicing strengths on five frequency bands.*

Comparing tendencies of significantly different measures on the AMI-IDIAP data and tendencies of the same measures on the NECA data (Table 2), we have found that voicing strengths and Fourier magnitudes have the same tendencies in both data bases. Most of these measures are also significantly different for the three classes (most, least dominant and neutral). The tendencies on Spectral entropy and Melcepstrum are not so clear. In Figure 2 means and quartile statistics of voicing strengths for male and female levels of dominance are presented. It is clear from this Figure that in general (without considering str1 for female, which was found to be not significant) voicing strengths separate quite well the two levels of dominance. The mean value of the neutral utterances is also shown, there is no clear tendency on neutral means in comparison with dominance.

In the following we checked how well the most dominant, least dominant and neutral classes discriminate with different sets of features. Table 3 shows classification results after a SFFS-LDA classifier for all, male and female data sets. In each case, and due to the very different distribution of the data (see Table 1), we have randomly selected M equal number of utterances for each class and performed the classification 20 times, the averaged classification rate is presented on the table. Additionally the LDA classification is N-fold cross validated. For male data M=100 and N=5, for female data M=60 and N=3, and for all the data M=200 and N=10. So every time the same number of utterances (20) is used for validation (testing) and the remaining data for training.

The discrimination results obtained with low level features (mcep + voiced/frame based) are comparable to the best results obtained for this data, in particular for male and all data. The features that better discriminate the data were voicing strengths and Melcepstrum.

Table 3: *AMI-IDIAP dominance: SFFS-LDA classification results for three utterance classes, most and least dominant and neutral. Voiced/frame based measures are low level acoustic measures (except Melcepstrum) where just voiced frames are averaged, that is frames for which $f0 > 0$.*

Measures	All	Male	Female
prosody	45.0	43.8	50.8
vq	41.9	44.2	48.9
vq + prosody	44.3	43.0	51.2
vq + prosody + mcep	50.5	56.2	61.7
vq + prosody + mag + mcep	51.8	56.2	51.6
voiced/frame based	47.2	50.9	52.6
mcep + voiced/frame based	57.8	55.1	55.4
prosody + voiced/frame based	49.8	47.7	48.4
vq + voiced/frame based	56.8	49.2	49.9
all features	55.4	56.1	58.1

6. Conclusions

Regarding the research questions addressed in this paper we can conclude that: (i) based on a comparison of vocal effort in controlled data and dominance in AMI-IDIAP meetings with the same acoustic measures, we have found that there exist a correlation between the tendencies of voicing strengths and Fourier magnitude measures of loud and soft vocal effort and most and least dominance. (ii) The discrimination of these dominance tendencies on the AMI-IDIAP meetings was quantified with a SFFS-LDA classifier, showing that in general the discrimination using low level features is around 55%. The results also show that other measures like prosody and voice quality can be also useful to discriminate dominance, but it is clear that not all the variability in dominance can be explained by acoustic features. This means that other cues like visual or textual might help to obtain a better classification as it was shown in [3, 2].

From the viewpoint of speech synthesis, the results are very important. First of all we have shown that low level acoustic features discriminate very well among different levels of vocal effort, so the effects can be modelled and controlled at this level. And secondly, we have probed that it is possible to detect dominance (and maybe other social effects) in more realistic data without the need of complex and difficult to calculate measures, like the ones used in [5]. What is clear, is that we need to combine low level acoustic features with other cues like visual or textual to obtain better prediction.

As future work, we will test at level of synthesis and perception, how much vocal effort can be controlled with variations on the low level features analysed in this paper.

7. Acknowledgements

This work is supported by the EU project SSPNet (FP7/2007-2013).

8. References

[1] J. E. Driskell, B. Olmstead, and E. Salas, “Task cues, dominance cues, and influence in task groups,” *J. Appl. Psych.*, vol. 78, no. 1, pp. 51–60, 1993.

[2] D. B. Jayagopi, S. Ba, J. Odobez, and D. Gatica-Perez, “Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues,” in *Proc. 10th Int. Conf. on Multimodal Interfaces*, Chania, Crete, Greece, 2008, pp. 45–52.

[3] H. Hung, D. Jayagopi, C. Yeo, G. Friedl, S. Ba, J. Odobez, K. Ramchandran, N. Mirghafori, and D. Gatica-Perez, “Using audio and video features to classify the most dominant person in a group meeting,” in *ACM Multimedia*, 2007, pp. 835–838.

[4] R. Rienks and D. Heylen, “Dominance detection in meetings using easily obtainable features,” in *Proc. Workshop on Machine Learning for Multimodal Interaction*, 2006, pp. 76–86.

[5] M. Charfuelan, M. Schröder, and I. Steiner, “Prosody and voice quality of vocal social signals: the case of dominance in scenario meetings,” in *Proc. Interspeech*, Makuhari, Japan, 2010.

[6] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2009. [Online]. Available: http://www.R-project.org

[7] M. Schröder and M. Grice, “Expressing vocal effort in concatenative synthesis,” in *Proc. 15th Internat. Congress of Phonetic Sciences (ICPhS)*, Barcelona, Spain, 2003.

[8] J. Carletta and et al., “The AMI meeting corpus: A pre-announcement,” in *Proc. Workshop on Machine Learning for Multimodal Interaction*, 2005, pp. 28–39.

[9] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, “Modeling dominance in group conversations using nonverbal activity cues,” *Trans. Audio, Speech and Lang. Proc.*, vol. 17, pp. 501–513, 2009.

[10] W. C. Chu, *Mixed excitation linear prediction*, ser. Speech coding algorithms Foundations and Evolution of Standardized Coders. Wiley, 2003, ch. 17, pp. 454–485.

[11] S. Imai, T. Kobayashi, K. Tokuda, T. Masuko, K. Koishida, S. Sako, and H. Zen, “Speech signal processing toolkit (SPTK), Version 3.4,” http://sp-tk.sourceforge.net, 2009.

[12] H. Misra, S. Ikbal, S. Sivadas, and H. Bourlard, “Multi-resolution spectral entropy feature for robust ASR,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.

[13] K. Sjölander, “The Snack sound toolkit,” http://www.speech.kth.se/snack, 2006, Swedish Royal Technical University (KTH).